УДК 004.85

Д.С. Сергеєв
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Україна
пр. Перемоги, 37, м. Київ, 03056

КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ КОГНІТИВНОГО АСПЕКТУ ОБРОБКИ ПРИРОДНОЇ МОВИ НА ОСНОВІ ПРИРОДНО-МОВНОЇ БАЗИ ЗНАНЬ

D.S. Sergeiev
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute», Ukraine
37, Peremohy av., Kyiv, 03056

COMPUTER MODELING OF THE COGNITIVE ASPECT OF NATURAL LANGUAGE PROCESSING BASED ON A NATURAL LANGUAGE KNOWLEDGE BASE

У даній роботі розглядається проблема взаємодії «людина-комп'ютер» у форматі дослідження засобів реалізації природно-мовної комунікації у технічних системах, зокрема проблеми розуміння природної мови. Пропонується підхід до вирішення цієї проблеми шляхом моделювання когнітивного аспекту мовленнєвої діяльності людини на основі природно-мовної бази знань. Оцінка моделі проводиться шляхом експериментальної перевірки наявності окремих ключових проявів, пов'язаних з розумінням тексту, в роботі бази знань.

Ключові слова: природна мова, розуміння природної мови, база знань, знання, квант знань.

This paper deals with human-computer interaction at the level of examination of methods of implementation of natural language communication in technical systems, in particular of natural language understanding. An approach to solving this problem is proposed which is centered around modeling the cognitive aspect of human speech activity on the basis of the natural language knowledge base. The model evaluation is conducted by executing an experimental test of occurrence of certain essential manifestations, linked to the understanding of natural language text, in the operation of knowledge base.

Keywords: natural language, natural language understanding, knowledge base, quantum of knowledge.

Вступ

Взаємодія людини з комп'ютером *(HCI, Human-Computer Interaction)* є однією з ключових проблем у галузі штучного інтелекту. Основною задачею, в рамках цієї проблеми, є трансформація сигналів, що надходять від людини-оператора, у зрозумілу комп'ютеру форму, та обернена трансформація результатів виконання запиту.

Хоча сучасні інтерфейси і способи взаємодії з комп'ютером активно розвиваються, основним інструментом для опису та вирішення складних задач є природна мова. Природна мова не завжди безпосередньо використовується як засіб комунікації, але у багатьох випадках інші типи інтерфейсів так чи інакше залежать саме від мови. Підписи на графічних іконках, мови програмування як спрощена демонстрація природної мови, символьні мови і т.д. часто зводяться до текстового опису відповідних функцій, операцій тощо. Фактично, уся проблема створення природно-мовного інтерфейсу тісно пов'язана з проблемою моделювання мовленнєвої діяльності людини взагалі – адже незалежно від вигляду та обсягу вхідного тексту, алгоритми його обробки залишаються ті ж самі.

Проблема розуміння комп'ютером природної мови на сьогоднішній день є актуальною. Обчислювальні потужності комп'ютерів протягом останніх десятиліть

зросли до фантастичних величин, але природно-мовні інтерфейси досі знаходяться на досить низькому рівні розвитку. При цьому, хоча люди добре володіють мовою з раннього віку, моделі мовленнєвої діяльності досі створено не було.

Постановка проблеми

Для вирішення проблеми розуміння тексту необхідно в загальному вигляді розв'язати дві основні задачі: отримання природно-мовного тексту у придатному для обробки комп'ютером вигляді та виділення власне знання, або смислового навантаження, з тексту.

Першу задачу в рамках даної статті не розглядаємо, оскільки вже існує багато сторонніх програмних засобів (наприклад, Stanford Parser [3]), здатних досить ефективно перетворювати аудіо-записи або графічні матеріалу у синтаксично розмічений фрагмент тексту. Крім того, введення тексту з клавіатури, хоча й не є оптимальним способом з точки зору швидкості роботи, має досить високі показники точності.

Натомість, задача, яка розглядається у статті – це власне задача виділення знань, тобто отримання з тексту знань у такому вигляді, який дозволяє ефективно обробляти їх математичними та логічними засобами.

Аналіз останніх досліджень і публікацій

Існує декілька основних підходів до виділення та зберігання знань.

Першим, найпростішим з точки зору реалізації, підходом до збереження знань є різноманітні електронні енциклопедії та довідники. Оскільки окремі елементи знань (статті) у таких системах пов'язані між собою мережею взаємних посилань, можемо виділити їх як окремий клас саме систем зберігання знань. Втім, можливість використання їх для безпосереднього виділення знань є сумнівною: найменший елемент – стаття – часто містить багато пов'язаних між собою фактів і тверджень, а наповнення таких систем знаннями та їх обробка в автоматичному режимі взагалі неможливі. Найвідоміші сучасні представники систем цього класу – Wolfram Alpha [6] та Google Knowledge Graph [4].

Більш формальним варіантом енциклопедій є фреймові системи збереження знань. Виділення знань у таких системах зводиться до заповнення відповідних полів фрейму необхідними даними. Хоча структура окремого фрейму може добре підходити для вирішення певної конкретної задачі, або навіть бути динамічною та мати деяку внутрішню логіку, вона все ж є досить строго описаною – а отже, не може одночасно охопити усі рівні абстракції. Як правило, ця структура або дуже складна сама по собі, що розширює сферу її використання, але робить логіку обробки знань дуже складною; або строго описана та жорстка, що призводить до втрати великої кількості знань при її наповненні та сильно обмежує можливі галузі її використання. Прикладом такої системи є семантичний *Web* за версією W3C [5].

Іншим популярним підходом є протилежний варіант – використання граматичної структури тексту, отриманої за допомогою синтаксичних аналізаторів, як семантичної його структури. Хоча використання існуючих аналізаторів позбавляє необхідності розробляти окремий програмний продукт для реалізації семантичного аналізу, проблеми синтаксичного аналізатора так само переходять на рівень семантики. Зокрема, усі результати роботи стохастичних алгоритмів синтаксичного аналізатора прямо переходять у БЗ, що зумовлює виникнення неточностей ще до початку власне процесу аналізу. Крім того, оскільки об'єкти «словосполучення» та «речення» досі не визначені, структура бази знань залишається в такому випадку незрозумілою. Прикладом такої системи є [1].

Проблему структури бази знань вирішує підхід використання семантичних мереж, окремим елементом знань у яких є слово. Ці системи добре підходять для вирішення цілої низки задач, але, на жаль, розуміння природної мови до них не входить: оскільки усі відношення між словами виносяться у «зв'язки» в рамках бази знань, обробка цих зв'язків стає окремою проблемою, особливо при необхідності виділити лише частину релевантних вузлів.

Виділимо спільні недоліки цих систем – основні проблеми виділення знань з тексту:
– не визначено окремий елемент знань: або це нестабільна структура (синтаксичні аналізатори), або негнучка (фреймові системи), або занадто велика (енциклопедії), або занадто мала (семантичні мережі);
– система або універсальна, або точна: враховуються лише найпопулярніші значення даного концепту або використовуються лише ті значення, які є несуперечливими у даній вузькій предметній області;
– у більшості сучасних автоматизованих систем виділення знань використовуються стохастичні підходи, що призводить до виникнення певної похибки ще до початку власне аналізу значення тексту.

Мета дослідження

Метою даної роботи є моделювання процесу розуміння природно-мовного тексту на основі природно-мовної бази знань, що характеризується використанням контексту та накопичених раніше знань при обробці нових вхідних текстів.

Основна частина

Об'єктом даного дослідження є природно-мовна база знань, розроблена автором на засадах інтеграційного підходу до моделювання мовленнєвої діяльності людини [2]. Головною особливістю цього підходу є визначення базової семантико-синтаксичної структури довільного природно-мовного повідомлення, що відповідає кванту знань – ситуації сенсорного рівня.

Використання кванту знань як базової семантико-синтаксичної структури дозволяє представити будь-який текст у вигляді сукупності таких структур та логічних відношень між ними, що, в свою чергу, значно полегшує його автоматичну обробку.

Предметом дослідження роботи є особливості описаної вище бази знань, зокрема можливість її використання для моделювання когнітивної (отримання нової інформації з власного досвіду) діяльності людини.

Основними структурними елементами бази знань є:
– структури *S*, що є формалізованим представленням базових семантико-синтаксичних структур – квантів знань. Кожна структура відповідає окремому кванту сенсорних знань або його уявному аналогу у випадку абстрактних ситуацій;
– відношення *R*, що поєднують окремі кванти знань. Відношення не мають реального сенсорного прототипу і є суто логічними структурами;
– маркери *M*, що пов'язують кванти знань та відповідні їм фрагменти тексту. Кожний маркер містить інформацію про зв'язок між словами та їх ролями у структурі, зв'язок між структурою та вхідним текстом, метадані щодо вхідного тексту тощо.

Кожна структура *S* складається з *Obj* (об'єкта), *Mov* (його дії) та їх атрибутивного оточення – *Attr(Obj), Attr(Mov), Attr(Attr(*))*. Кожне відношення поєднує 2 окремі структури. Таким чином, загальна структура довільного фрагменту знань – від одного кванту знань до усього обсягу бази знань – може бути описана сукупністю ситуацій *S* та відношень *R*.

Наповнення бази знань

Поставлена мета потребує демонстрації особливостей роботи природно-мовної бази знань. Оскільки найбільш логічним способом їх перевірки є перевірка виконання певних запитів, надалі концентруємо увагу саме на них. Процес наповнення БЗ відповідними знаннями описуємо лише коротко.

Заради чистоти експерименту використовуємо примітивний лінгвістичний процесор та враховуємо лише один тип відношення – симетричне відношення «тире» або «є»: «S_1 – S_2» ; «S_1 є S_2». Як вхідні тексти використовуємо такі, що не потребують попереднього знання у галузі: підручник «Природознавство» для 4 класу за авторством Т.Г. Гільберта та підручник «Астрономія» для 11 класу за авторством М.П. Пришляка, тема *«Космос»*.

Об'єктом дослідження обираємо основний концепт *«планета»* і його похідні. Концепт у даному випадку відображає фрагмент знань, над яким виконується експеримент: він може містити від одного слова до повної структури *S*, або навіть декілька структур, пов'язаних відношеннями. У даному випадку використання терміну «слово» обмежує варіативність лише однією словоформою, терміну «лексема» – концептами, які можна описати лише одним словом.

Підручник «Природознавство» містить загальні відомості про планети: *«Земля – наша планета», «Нептун – планета», «перша планета», «найхолодніша планета»*, тощо. Підручник «Астрономія» містить знання більш наукового спрямування: *«планета – космічне тіло», «планета обертається», «планети рухаються»* і так далі. Отже, після початкового наповнення отримуємо БЗ, що містить знання у вигляді окремих *S*, кожну з яких можемо розглядати як окрему сутність, в основі якої завжди є концепт – *Obj*. У маркерах, що відповідають цим *S*, зберігається інформація про джерело тексту, в даному випадку – назва підручника. Звісно, це може бути будь-яка інформація – і отримана автоматично з метаданих, і додана іншими методами тощо.

Семантика концепту

У першу чергу перевіряємо можливість автоматичного виділення концепту та формування його семантики. Оскільки текстове представлення концепту відоме, задача його виділення з власне тексту повністю виконується лінгвістичним процесором. Отже, оскільки БЗ не має впливу на результат цієї дії, його не перевіряємо.

Зазначимо, що концепт може описуватись і цілою структурою з кількох квантів та відношень між ними – але в такому випадку цей концепт можна розбити на складові кванти знань, що є окремими концептами, пов'язаними логічними зв'язками *R*. Отже, перевіряти роботу БЗ у такому випадку нема необхідності. Розглянемо формування семантики концепту на прикладі згаданого вище концепту *«планета»*.

Виділення знань по запиту *«планета»* дозволяє визначити його атрибутивне оточення (*«найгарячіша»*, *«найхолодніша»*, *«перша»*, …), можливі дії (*«рухається»*, *«обертається»*, …) та зовнішні відношення (*«Земля – планета»*, *«Марс – планета»*, …). Таким чином, БЗ дозволяє отримати повне семантичне оточення слова (з або без урахування маркерів). На цьому етапі відмінність від класичних систем, зокрема семантичних мереж, майже відсутня: побудова мережі сусідів даного слова є досить простою задачею.

Розширимо концепт від одного слова *«планета»* до словосполучення (*Obj* + *Attr*) *«перша планета»*. Для отримання семантичного оточення для *«перша планета»* достатньо вибрати з БЗ усі випадки, коли ці слова пов'язані у кванті знань як *Obj* та

Attr відповідно. Результат пошуку за цим запитом містить набагато менше зв’язків, відношень, а отже – і результатів.

Одразу відзначимо першу особливість природно-мовної бази знань. У рамках окремого кванту знань ролі і зв’язки між ними визначаються взагалі досить просто; при введенні зовнішніх відношень складність задачі не зростає кардинально; і, як було показано вище, при збільшенні обсягу тексту структура знань залишається такою ж самою. Це дозволяє зробити перший важливий висновок:

В.1.: *виділення окремих квантів знань та відношень між ними відбувається автоматично;*

Крім того, архітектура бази знань передбачає зберігання будь-якого формату природно-мовного тексту, а отже:

В.2.: *точність та повнота обробки тексту залежать лише від лінгвістичного процесора та модуля відношень.*

Моделювання розуміння тексту

Досить часто концепти можуть бути представлені рівноправними синонімами – словами або фрагментами тексту. Так, з фрагменту тексту *«перша, найменша планета»* отримуємо знання про те, що *«перша планета»* також є *«найменшою»* – але це не розповсюджується на інші планети. Звісно, іноді такі синоніми будуть нерівноправними, або ж неоднозначними – але вирішення цих питань можемо покласти на вже безпосередньо системи обробки знань, оскільки логіка відношень є окремою темою для дослідження. У будь-якому випадку, виникає проблема згортки знань – адже різні фрагменти знань, що представляють один і той самий концепт, мають різне семантичне оточення.

Таким чином, приходимо до моделювання узагальнення – тобто обміну семантичним оточенням між синонімами. При виникненні відношень певного типу (наприклад, симетричного *«є»*) між окремими фрагментами знань, необхідно лише взаємно провести нові зв’язки і відношення між відповідними квантами знань. Це досить проста операція, але на великих обсягах даних вона може бути затратною по ресурсах, тому не виносимо її в обов’язкові функції БЗ. Отже:

В.3.: *згортка семантичного оточення синонімів може відбуватись автоматично;*

Очевидно, що при виконанні згортки можуть виникнути нові потенційні зв’язки між елементами семантичного оточення первинних синонімів, які теж можна розкрити. Це означає, що:

В.4.: *згортка семантичного оточення синонімів може мати рекурсивний характер.*

Єдине, що залишилось змоделювати – використання описаних вище принципів при внесенні нових знань у БЗ. Оскільки немає суттєвої різниці, звідки було отримано нові знання – з зовнішнього світу або з внутрішньої роботи БЗ – можемо стверджувати, що процес обробки буде проходити за тією ж самою схемою.

Зазначимо окремо, що отримана модель досить точно описує процес не лише розуміння, але й набуття нових знань людиною, оскільки при оновленні фрагменту знань автоматично виникає потенціал для оновлення у всьому його семантичному полі. Більш того, оскільки ресурси комп’ютера, як і людини, не є безмежними – цілком імовірне виникнення моделі цього процесу з визначення пріоритету, затримки виконання, періодичного оновлення знань тощо – тих самих процесів, які спостерігаємо у вищій нервовій системі діяльності людини.

Висновки

Використання природно-мовної бази знань для моделювання когнітивного аспекту мовленнєвої діяльності людини можемо вважати успішним.

Автоматичне виділення окремих квантів знань та відношень дозволяє заповнювати БЗ знаннями з довільних текстів, причому якість цього процесу обмежена лише якістю роботи лінгвістичного процесора.

База знань у нормальному режимі роботи дозволяє змоделювати розуміння комп'ютером природно-мовного тексту: узагальнення подібних концептів, визначення окремих слів і квантів знань та їх семантики, врахування контексту запиту (на прикладі джерела тексту) та маркерів текстів у БЗ. Оновлення семантичних зв'язків при доповненні бази знань відбувається за схемою, подібною до схеми осмислення нового тексту людиною.

Отримані з тексту природно-мовні знання мають чітко визначну структуру, що дозволяє передавати їх для подальшої обробки у системи обробки знань або вдосконалювати саму ПМБЗ відповідним чином для роботи з ними.

Використання описаного вище підходу дозволяє автоматизувати наповнення довільної бази знань з матеріалів природно-мовного тексту, що може бути використано фактично в усіх сучасних інформаційних природно-мовних технологіях.

Література

1. Björkelund A. A high-performance syntactic and semantic dependency parser / A. Björkelund, B. Bohnet, L. Hafdell, P. Nugues // Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (COLING 2010). — 2010. — No. August. — P. 33–36.
2. Kyslenko Y. Cognitive architecture of speech activity and modelling thereof / Y. Kyslenko, D. Sergeiev // Biologically Inspired Cognitive Architectures. — 2015. — Vol. 12.
3. Marneffe M.-C. De Generating typed dependency parses from phrase structure parses / M.-C. De Marneffe, B. MacCartney, C.D. Manning // Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). — 2006. — P. 449–454.
4. Singhal A. Official Google Blog: Introducing the Knowledge Graph: things, not strings/ A. Singhal. — 2012.
5. W3.org Semantic Web - W3C / W3.org. — 2012.
6. Wolfram|Alpha Frequently Asked Questions.

RESUME

D.S. Sergeiev

Computer modeling of the cognitive aspect of natural language processing based on a natural language knowledge base

This deals with human-computer interaction at the level of examination of methods of implementation of natural language communication in technical systems, in particular of natural language understanding. The main objective of this article is modelling of certain aspects of human speech activity, namely the process of understanding of given natural language text, by using a natural language knowledge base.

The first part of the paper analyses different approaches to retrieving and storing of natural-language information, highlights the pros and cons of every approach and shows their common problems. The paper concludes that encyclopedic knowledge bases are not suited for use in automatic knowledge processing systems, frame-based approaches limit the integrity and consistency of natural language information and semantic networks generally have insufficient systems of relation between nodes. Lack of common robust structure of natural language text is established as a common flow of the examined systems.

The second part explains the architecture of a natural-language knowledge base and its components, as well as the process of filling the knowledge base and preparing it for the experiment. The basic semantic-syntactic structure, the cornerstone entity of the integrated approach to modelling of speech activity, is proposed as the basis for the natural language

knowledge base. The structure of such knowledge base is described as set of such structures S, relations between them R and semantic markers M. Accordingly, the natural language text contains individual concepts, represented by mono-predicate structures and relations that connect these structures into poly-predicate structures. Each mono- or poly-predicate structure is shown to have a corresponding marker that link every element of the said structure to a certain natural language construct such as a word. The marker is also designed to store information about the original text in order to allow restoring the text from the given part of the knowledge base.

The last part outlines the process of adding new knowledge to the knowledge base and describes how, with regard to the limitations of the model of knowledge base, relations and markers for text fragments are formed and maintained. The consequent process of formation of semantic context of a given structure is analyzed. In particular, the analyzed phenomena include formation of semantic context of given structure from a single fragment of text, expanding of semantic context of given structure from several conjoined text fragments and automated formation of several independent context objects from different text fragments based around the same structure.

The article concludes that retrieving knowledge from natural language text by making use of the knowledge base to allows to emulate natural language understanding with sufficient quality of the emulation, the main limitation being the capabilities of the linguistic processor module.

Надійшла до редакції 29.11.2016